## TOKENIZER FOR A NATURAL LANGUAGE PROCESSING SYSTEM

## ABSTRACT OF THE DISCLOSURE

The present invention is a segmenter used 5 natural language processing system. The in segmenter segments a textual input string into tokens language processing. further natural for accordance with one feature of the invention, segmenter includes a tokenizer engine that proposes 10 linguistic and submits them to а segmentations knowledge component for validation. In accordance of invention, the another feature the with segmentation system includes language-specific data that contains a precedence hierarchy for punctuation. 15 input string contain proposed tokens in the punctuation, they can illustratively be broken into subtokens based on the precedence hierarchy.